Applying the Methodology of Learner Corpus Analysis to Telecollaborative Discourse

NINA VYATKINA

Telecollaborative pedagogy and the research methodology of learner corpus analysis are both relative newcomers on the foreign language education scene. The aim of this chapter is to show how corpus analysis methods and tools can be applied by researchers interested in examining telecollaborative discourse. I will show how the two have been successfully brought together in existing research and suggest directions for future applications. The chapter begins by describing learner corpora and the research method of contrastive interlanguage analysis. In particular, I define the terminology used in this research paradigm, describe its characteristics and purposes, and list the contexts where it has been applied. The next section shows why this methodology is uniquely suited to research on intercultural computer-mediated exchanges. In what follows, I provide concrete examples of how selected corpus research methods and tools have been applied in collecting and analyzing a telecollaborative corpus in research exploring language development in college-level learners of German. I will conclude by a discussion of results and suggestions for future research.

Language corpora and Contrastive Interlanguage Analysis (CIA)

This method has been used in both descriptive and contrastive language studies to explore variation across genres and registers – a research area made famous by Biber (1988) and Sinclair (1991). Most corpus-based studies have focused on distribution and frequency of linguistic features as used by Native Speakers (NSs) of different languages (although predominantly English). Findings from this research have attracted attention of Second Language Acquisition (SLA) and Foreign Language Teaching (FLT) researchers who expanded corpus-based research to include not only NS data but also learner data. For this purpose, learner productions are systematically collected and compiled in electronic *learner corpora*:

> Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogenous way and documented as to their origin and provenance. (Granger, 2002: 7)

The research method based on comparisons between NS and learner corpora has been termed *Contrastive Interlanguage Analysis*, or *CIA* (Granger, 1998a). According to Granger, CIA studies seek to uncover "qualitative differences (misuse) and quantitative differences (over- and underuse)" of the targeted features by learners in comparison to NSs (2003: 541). Although learner corpora are relative newcomers (see Pravec, 2002; Nesselhauf, 2004, for reviews) and CIA studies span only about a decade, they have greatly enriched SLA research with findings grounded on large amounts of empirical data. Areas of inquiry have ranged from morpho-syntactic constructions to lexis to pragmatic features and discourse patterns (see collections by Granger, Hung, & Petch-Tyson, 2002; Aston, Bernardini, & Stewart, 2004; Sinclair 2004; Aijmer, 2009). CIA has proven especially productive in revealing subtle differences between learners at higher proficiency levels (high intermediate to advanced) and NSs. Consid-

erable advances have been made by researchers working on the collection and analysis of the International Corpus of Learner English (ICLE)[1] and the Louvain International Database of Spoken English (LINDSEI)[2]. Working with a number of subcorpora encompassing the written and oral productions of English as a Second Language (ESL) learners with different first language (L1) backgrounds, these researchers have compared learner and NS use of English. For example, Granger (1998b) found that learners overuse generic adverbs and adjectives (such as *very* or *important*) but under-use their more specific and contextually restricted counterparts (such as *closely, highly* and *critical, significant*). Aijmer (2002) showed that learners used significantly more modal verbs in their writing than NSs, which contributed to an overuse of spoken patterns inappropriate in a written academic genre. Nesselhauf (2005), Gilquin, Granger, & Paquot (2007), Gilquin (2008), and Paquot (2008) explored differences in NS and learner use of phraseological expressions such as noun-verb combinations (e.g. *to have a look*) and discourse markers (e.g. *on the contrary* in writing and *sort of* in speaking). These studies again show that advanced learners' productions, without being grammatically inaccurate, differ from baseline NS writing in stylistic nuances and idiomaticity. Rather than using the terms 'overuse' and 'underuse' as disparaging labels for learner deficiencies, the above studies aim at raising awareness of learners, teachers, and researchers of second language (L2) "Nuancenkompetenz", or 'nuance competence' (Weinrich, 1993: 842). Notably, this research has resulted in a number of corpus-based pedagogical suggestions in the areas of pedagogical lexicography (Rundell & Granger, 2007) and phraseology (Granger & Meunier, 2008).

This brief overview highlights advances made in learner corpus research to date. However, much of the potential of computer-aided corpus analysis remains underexplored due to a number of limitations of existing learner corpora. First, most of them are monolin-

1 <http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm>.

2 <http://cecl.fltr.ucl.ac.be/Cecl-Projects/Lindsei/lindsei.htm>.

gual, i.e. researchers require an *external* NS comparison corpus in order to conduct contrastive learner corpus analyses. However, the explanatory power of such comparisons is limited because, as Vyatkina & Belz (2006) argue, "the data to which learner productions are compared were produced at a different point in time, under different circumstances, and in different contexts" (p. 320). Second, with some notable exceptions, learner corpora are highly restricted with respect to genre (most of them include only written argumentative essays) and language (predominantly L2 English). Finally, the majority of existing corpora are cross-sectional and synchronous. They comprise data collected as snapshots of performance of particular learner populations at one point in time. As Belz points out, studies based on such learner corpora "tend to describe L2 use only at a particular point in time. As a result, they are not positioned well to address questions at the heart of SLA research, namely, how does L2 competence *change* over time?" (2007: 47).

Some researchers make inferences about typical developmental stages and patterns based on corpus data collected from different groups of learners at different stages in collegiate instructional sequences (Tono, 2000) or of different age (Housen, 2002). However, as Housen (2002) notes, the aggregate view of this largely cross-sectional learner corpus research may mask variation in individual developmental paths taken by learners. Yet, despite "nearly ritualized" calls (Ortega & Byrnes, 2008: 5) for longitudinal SLA studies, dynamic accounts of variability in learner productions, more exhaustive sampling, and fine grained analyses of learner language (Ellis & Larsen-Freeman, 2006; Norris & Ortega, 2009), researchers largely shy away from the difficult and time consuming endeavor of compilation and analysis of longitudinal learner corpora. Examples of notable exceptions are longitudinal corpora of learner written German (Lüdeling, Walter, Kroymann, & Adolphs, 2005; Byrnes & Sinicrope, 2008) and learner spoken French (Myles, 2008) as well as the Lab School multimedia corpus containing textual, audio, and video data (Reder, Harris, & Setzler, 2003). In the following section, I discuss

how the compilation and analysis of *telecollaborative (TC) learner corpora* address the research needs and design limitations mentioned above.

Advantages of TC Data for Learner Corpus Research

Telecollaboration is one of the forms of social networking that has been on the rise in FLT during the last decade (see Warschauer, 1996; Furstenberg, Levet, English, & Mallet, 2001; Belz & Thorne, 2006; O'Dowd, 2007; Dooly, 2008; Guth & Helm, 2010; Dooly & O'Dowd, this volume). Belz defines telecollaboration as:

> institutionalized, electronically mediated intercultural communication under the guidance of a languacultural expert (i.e., a teacher) for the purposes of foreign language learning and the development of intercultural competence. (2003a: 2)

One unique feature of TC classes is "the alternation of Internet-mediated *inter*cultural sessions with face-to-face *intra*cultural sessions" (Belz, 2006: 214, emphasis in original). During traditional classroom sessions, the teacher provides instruction in various L2 grammatical, pragmatic, and cultural topics, while during computer-mediated communication (CMC) sessions, learners practice their acquired knowledge in real-life interactions with native speakers. During subsequent instruction sessions, the teacher gives the students an opportunity to observe their own accurate and inaccurate L2 uses in excerpts from their CMC discourse as well as to discuss it *in plenum* or individually. The instructor teacher thus acts as a "facilitator" (Rogers, 1969: 105–106) who provides guidance to learners during both intercultural and intracultural sessions. Moreover, the teacher in TC courses also is in a unique position to conduct classroom-based research using corpus analysis methods. Features that make TC data especially suitable for learner corpus construction and CIA studies are addressed in the next sections.

Automatically Saved Data

Because TC language practice sessions are always computer-mediated and thus L2 productions are usually automatically saved, this presents a unique opportunity for corpus data collection. This holds true for text-based CMC genres such as email, forum, and chat as well as audio/video-based genres such as teleconferences. Provided the teacher/researcher follows ethical standards for research and an approved procedure for collecting human subject research data[3], these computer recordings of learner and NS productions can be used for construction of corpora for concurrent or future investigations. Whereas audio and video data would require subsequent transcription, text-based data obviate the need for this time-consuming part of research and can be directly transferred into an electronic database. This immediate availability of data for research greatly enhances the feasibility of the study, especially if the researcher has limited funding and is not part of a larger collaborative team.

Richness of Metadata

Smaller, custom-made corpora (Ghadessy, Henry, & Roseberry, 2001) usually contain rich participant metadata (on age, gender, language learning history) and task metadata (on topic, prompts, planning time). As Granger argues:

> the usefulness of a learner corpus is directly proportional to the care that has been exerted in controlling and encoding [learner and task] variables. (2002: 9).

3 Researchers need to plan well in advance and keep in mind that their research protocol must be approved by the Institutional Review Board (IRB) *prior* to the beginning of the data collection process, i.e., *prior* to the TC exchanges. Mackey & Gass (2005: 25–42) provide a very helpful guide for young researchers which discusses ethical issues in research involving humans, describes the process of obtaining informed consent, and gives helpful examples of necessary documents.

Additionally, such smaller specialized corpora can be organized in accordance with the teacher/researcher's own scholarly and pedagogical needs without compromising generally accepted standards for corpus collection (Seidlhofer, 2002; Flowerdew, 2005; Lee & Swales, 2006). This is especially true for TC corpora because the teacher/researcher has an insider view on the study context, which greatly enriches the ethnographic dimension (Flowerdew, 2005) and enhances the ecological validity of both the learner corpus itself and studies based on its data (Reinhardt, this volume).

Integrated NS Baseline Data

Tasks that teachers formulate for participants of TC exchanges (who are speakers of different languages) may include various configurations of languages to be used: 1) both partner classes continuously alternate the use of their respective L1 and L2 thus having an opportunity to practice their L2 but also an obligation to provide models in their L1 to their partners for whom it is an L2 (Belz, 2005a); 2) only one language is used, as L1 by one partner group playing the role of tutors and as L2 by another partner group playing the role of learners (Sauro, 2009); 3) only the L1 is used by both partner groups during intercultural (CMC) sessions, whereas L2 is practiced during intracultural (traditional, in-class) sessions (Furstenberg et al., 2001). Thus, in most TC courses, data in one and the same language are being produced by both NSs and learners of that language during the same or similar tasks. This unique feature of telecollaboration makes it a perfect locus to collect a learner corpus with an in-built NS comparison corpus. As Kasper and Rose argue, a "defensible standard" against which learner language can be measured should be "derived from successful multilingual speakers' interactions in activities relevant for a given learner population" (2002: 86). Following this line of argument, an integrated NS corpus represents a more defensible baseline for measuring learner productions than external NS corpora typically employed in CIA studies because all TC participants are multilingual

speakers participating in activities that are relevant to them. The validity of the NS baseline is the highest in configuration 1 (alternating L1 and L2), as Belz and Vyatkina argue, "[t]his configuration is an advantage in contrastive interlanguage analysis because learner productions are not separated in time and space from the baseline" (2008: 45). Moreover, members of both populations in configuration 1 participate in status-equal encounters (unlike in configuration 2) and not only in similar but also in the *very same* interactions (unlike in configuration 3).

Dense developmental data

Since TC courses typically span weeks or months of CMC exchanges, corpora comprising TC data represent a subtype of diachronic, or longitudinal, corpora, which allow tracking of participant development over time. Belz and Vyatkina propose density of observation as the main feature distinguishing the subtype of *Developmental Learner Corpora* (*DLC*) from other longitudinal corpora:

> We distinguish between *longitudinal* and *developmental* analyses with regard to the density of observation of learner performance over time. While 'longitudinal' may refer to analyses in which waves of data are elicited at more distant intervals (e.g. at the beginning and end of a semester), we reserve the term 'developmental' for those analyses in which learner performance is documented at close intervals or at all points of production. (Belz & Vyatkina, 2008: 33)

All TC data are saved in computers simultaneously and automatically, which means that researchers have access to the full contingent of learners' production data for the duration of the intercultural exchanges. Such access is especially valuable for developmental studies because no intermediate stages remain unaccounted as opposed to cross-sectional designs employing only two or three data elicitation waves. In other words, a TC researcher may collect and archive every single learner and NS utterance that each participant produces over the course of the TC partnership under study. This allows analysis of

development of each individual in reference not only to the NS baseline but also to previous performance of the same individual. Thus, a DLC lends itself not only to cross-sectional aggregate analyses (typical for CIA studies) but also to analyses in sequence, tracking microchanges of each focal participant over time (Belz 2003b; Belz & Kinginger, 2003; Kinginger & Belz, 2005; Vyatkina & Belz, 2006; Vyatkina, 2007).

Pragmatic and Interpersonal Linguistic Features

Another major advantage of TC corpora is the wide array of discourse types and associated linguistic features represented in the data. Notably, CMC discourse belongs to hybrid linguistic genres such as email and Internet relay chat that are "culturally conditioned on a cline of 'writtenness' and 'spokenness'" (McCarthy, 1993: 171). Kern supports this claim, positing that synchronous CMC "combines the temporal immediacy of spoken interaction […] with the social distancing allowed for by writing", and "incorporates many features of spoken mode within a written medium" (2000: 238).

Indeed, previous research has shown that TC discourse is rich in linguistic features typical of both spoken registers (e.g. personal pronouns, see Belz & Kinginger, 2003) and written registers (e.g. pronominal adverbs, see Belz, 2005b).

Due to the oral mode quality of the CMC data captured in the written electronic medium, TC corpora lend themselves to the exploration of *interpersonal language functions*, especially in teacher-guided project-based intercultural exchanges. TC partners need to express these functions if they participate in discussions about a wide variety of contemporary topics (Belz, 2005a). Although such TC data are produced in response to classroom tasks, they nevertheless come from real-life discussions on topics collaboratively chosen by the partners themselves and thus represent authentic interactional data. On the one hand, learners are requested to complete various communicative learning tasks by interacting with NSs including getting to

know each other, exchanging their opinions on assigned topics, comparing attitudes and viewpoints and the like. On the other hand, learners inevitably engage in a variety of additional communicative actions beyond classroom tasks such as personal relationship building, intercultural misunderstandings, delegation of project tasks, directions for the use of project mediating software, disagreements, and discussion of popular culture topics. TC communicative actions that have been researched (see Belz, 2007 for a review) range from language play (Belz & Reinhardt, 2004) and flirting (Belz & Kinginger, 2002) to crying (O'Dowd, 2006) to conflict (Belz, 2003b; Schneider & von der Emde, 2006). Participating in these communicative actions requires a broad array of interpersonal linguistic features (Biber, 1988). Therefore, TC corpora are uniquely conducive to exploration of development of L2 pragmatic competence as opposed to the majority of learner corpora, which are comprised of student writing in the register of monologic argumentative prose (see Pravec, 2002 for a review).

In what follows, I describe the procedure of creating one particular TC learner corpus and demonstrate what research methods have been used to perform corpus-based analyses on the material of this corpus. These specific examples are also intended to illustrate how the advantages of TC learner corpora described above play out in actual research.

Analyses of TC discourse based on *Telekorp*

The Corpus

The *Telecollaborative Learner Corpus of English and German*, or *Telekorp* (Belz, 2005a), contains all of the CMC discourse and associated metadata produced through six iterations of TC partnerships conducted between a large public US university and a German teachers' college.

In these TC courses, American undergraduate students who learned German as an L2 were paired electronically with German university students who studied English as an L2. In a series of teacher guided tasks, the partner groups communicated with each other about various intercultural topics via email and live chat using German half of the time and English the other half (see Belz, 2005a for a detailed description of the participants and the course). In terms of learner corpus classification, *Telekorp* represents a bilingual developmental learner corpus with an integrated NS comparison corpus. Because both transatlantic partner groups (Americans and Germans) used both languages alternately, *Telekorp* contains, in fact, two integrated NS-learner subcorpora: L2 German with an L1 German comparison baseline and L2 English with an L1 English comparison baseline.

The Database

The TC electronic textual exchanges were initially saved by the participants themselves in the teleconferencing client Open Text FirstClass®. On a weekly basis, a research assistant copied all production data, replaced all participant names with unique pseudonyms to protect their confidentiality, and entered them into an electronic database using FileMaker Pro® software (Fig. 1).

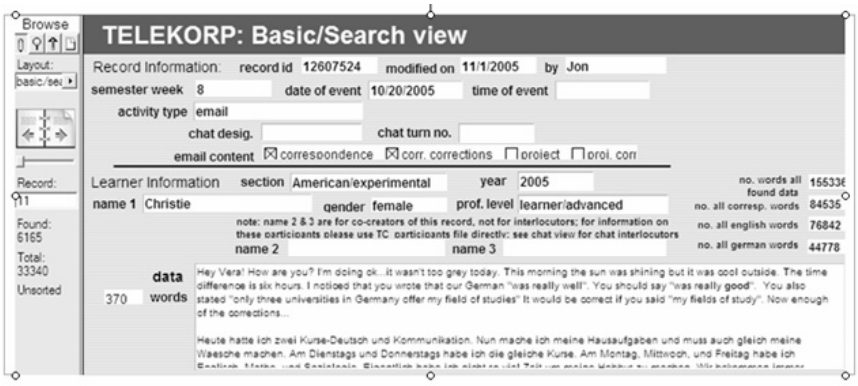

Figure 1. Example record of the *Telekorp* database

The full *Telekorp* database consists of three relational databases: 'Basic/Search View', 'Participants', and 'Full Chats'. More specifically, the production data were entered into the "data" field of the "basic" database and supplemented with a range of metadata: semester week and date of production; activity type (email, chat, survey, etc.); correspondence type (general correspondence, project discussions, error corrections, etc.); and a unique pseudonym of the author (or authors if the entry was written collaboratively). Each email constitutes one record for emails, and each turn produced by each participant constitutes one record for chats. All CMC data were further separated into the fields 'English words' and 'German words' to enable separate frequency counts for each language.

The basic view database is automatically linked to two relational databases: 'participants' and 'full chats'. The participant database is linked to the pseudonym of the first author (the person who typed the respective CMC entry) and allows for automatic display of the following learner metadata: year of participation in the TC course, section (e.g. American/ experimental), gender, and proficiency level. In sum, metadata for about 30 learner and task variables were collected at the beginning of the course via an electronic questionnaire and entered into *Telekorp*. The third relational database is linked to specific chat turn entries in the basic view and displays full chat records with sequential turns taken by all chat participants. FileMaker performs automatic record count and raw word (token) count in each field and record. Additionally, the corpus was supplemented by (non-CMC) production data (learners' written portfolio entries and oral interview transcripts), biographical survey data as well as researcher and instructor's field notes.

Target Linguistic Feature: German Modal Particles

A series of studies have explored the use of German Modal Particles (MPs) by learners in comparison to NSs in *Telekorp* as well as the effect of pedagogical interventions for teaching MPs using the same

corpus. MPs function as important interpersonal markers in NS German speech but are notoriously difficult for L1 English learners because of the absence of a direct translation into English, their rampant polysemy, and strongly context-bound meaning (König & Requardt, 1991; Möllering, 2004; Weydt 2006). Based on an exploratory analysis of *Telekorp*, Vyatkina (2007) ascertained that NSs regularly used MPs in CMC exchanges conducted in German. Furthermore, she found that 97–98% of MPs in the corpus were used by the NSs and only 2–3% by the learners. This gave the researcher a rationale to develop and administer a pedagogical intervention for teaching four German MPs (*ja, denn, doch, mal*) to the American students within the framework of two consecutive TC courses using *Telekorp*-based materials. The effects of these interventions were explored in a series of studies.[4] This chapter is based on Vyatkina's (2007) unpublished study and presents (in a slightly modified format) the data coding procedure, corpus analysis tools and methods, and selected results focusing on CIA methodology. The study sought to answer the following research questions: Will the learners use more MPs after the intervention? Will the learner use be NS-like as to the frequency and distribution across the CMC media (chat and email) after the intervention?

Design

Seven American intermediate learners of German and sixteen native speakers participated in the focal TC course and exchanged emails and chats during eight weeks in fall 2005. This exchange resulted in 330 emails (33,000 German words) and 30 chat sessions consisting of

4 The effects of the first intervention were described in Belz & Vyatkina (2005) and Vyatkina & Belz (2006). The design was subsequently fine-tuned for the next iteration of the study which, after completion, presented rich material for Vyatkina's (2007) dissertation. Some effects of the second intervention were reported in Belz & Vyatkina (2008) but most material remained largely unpublished.

5,800 turns (5,300 German words). Weeks 1–4 of the exchange constituted the pre-intervention condition of the study, and weeks 5–8 the intervention/post-intervention condition.[5] The pedagogical intervention for MPs was made up of three week-long cycles each including an explicit data-driven instruction session based on the quantitative and qualitative analysis of data produced by the participants during CMC practice sessions (see Fig. 2). The instructional sessions included teacher explanations and class discussions of the MP meaning, functions, and use supported by projecting TC corpus excerpts containing MPs on the big screen, handouts, and worksheets (see Vyatkina & Johnson, 2007).

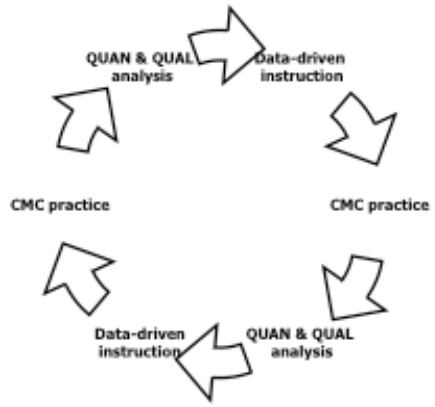

Figure 2. Cyclical pedagogical intervention

Data Coding

Textual data contained in *Telekorp* are raw, i.e. they are not coded for any linguistic categories such as parts-of-speech or syntactic struc-

5 One instructional session based on implicit enhanced condition was administered in week 4. However, it did not have any immediate production effect and is not considered here a part of the intervention (see Belz & Vyatkina, 2008 for details).

tures. Therefore, target linguistic features had to be manually tagged. McEnery, Xiao and Tono describe this "problem-oriented corpus annotation" as follows:

> First, it is not exhaustive –only the phenomenon directly relevant to a particular research question, rather than the entire contents of a corpus, is annotated. Second, the scheme for problem-oriented annotation is developed not for its broad coverage and consensus-based theory neutrality but for its relevance to the specific research question. (McEnery, Xiao and Tono, 2006: 43)

For MP coding, the *FilemakerPro* 'Find' tool was used to find each instance of the four focal words (*ja, mal, denn, doch*). Next, the word class membership of each found word was manually disambiguated to "weed out" (Poos & Simpson, 2002: 8) the homonyms with non-MP meaning. For example, the word *ja* can function in German as a modal particle and as an answering particle ('yes'), i.e. an MP homonym. As pointed out earlier, the MPs cannot be directly translated into English and require a paraphrastic description. For example, Möllering explains the meaning of the MP *ja* as follows:

> the speaker is trying to establish 'common ground' by marking a proposition as known to the hearer, thus inviting him/her to either accept the proposition as premise for the following exchange or to ask for clarification. (2004: 237)

For *Telekorp* coding purposes, the MPs were assigned the index 1 and MP homonyms the index 2, immediately following the word (e.g. *ja1; mal2*). For example:

> Das ist *ja1* cool!! 'This is [I am certain you agree] cool!!'

> Paula: ist es schweirig? 'is it difficult?'
> Simone: *ja2* 'yes'

This manual coding was the necessary initial step for automated frequency analyses performed later only on target words coded with the index 1.

Wordsmith Tools

The general learner corpus analysis technique is, according to Barlow:

> trawling through learner corpora using searching software to reveal and quantify recurrent patterns, typically lexico-grammatical patterns, that characterize the learner language associated with different learners and different settings. (2005: 336)

For MP frequency comparisons, the lexical analysis software *WordSmith Tools*®[6] (Scott, 2008) was used. It allows for retrieving examples and comparing frequencies of occurrence of focal features and their co-occurrence with other linguistic elements in specified subsets of data. Furthermore, the software automatically provides basic statistical indicators for comparing datasets such as log-likelihood index or type-token ratio. In addition to frequency, *WordSmith* provides information about, and illustrations of, the distribution of focal elements in the corpus subsets.

As *WordSmith* works with plain text files, data subsets intended for analysis were first exported from *FileMaker* into Unicode text files (to preserve special German typographic characters such as umlauts). In particular, data subsets for all learner productions before the intervention, all learner productions after the intervention, and all NS productions were extracted. For some analyses, these subsets were further subdivided into email data and chat data.

The three main tools provided by *WordSmith* are *Concord, KeyWords*, and *WordList*. The following sections describe selected comparative analyses performed with these tools on *Telekorp* MP data for the 2005 participant cohort (for the full reporting on the analyses, see Vyatkina, 2007).

6 <http://www.lexically.net/downloads/version5/html/index.html>.

Dispersion

In order to acquire an overall impression of the distribution of the target features in the corpus, the dispersion plots provided by the *WordSmith Concord* tool are useful. Dispersion plots are "maps showing where in the texts the search words were found" (Scott, 2001: 47). For this study, dispersion plots for each focal MP and each student population (learners and NSs) were retrieved from *Telekorp*. In both plots shown in Figure 3, each vertical line represents one MP use. The left and right margins represent the beginning and the end of the TC Internet exchange (spanning the total of nine weeks). The comparison of the two distribution plots for NS and learner MP use makes visible at what point the pedagogical intervention was delivered. NSs use all four focal MPs relatively evenly, with somewhat higher density of the MP *ja* and an increase in the overall density in the final third of the course. In contrast, three single lines in the left half of the learner plot represent three single MP uses in the first half of the TC exchange. The right third of the learner plot demonstrates a sharp increase (corresponding to Week 6 of the exchange, i.e., a week after the beginning of explicit instruction in Week 5) with the highest density of use for the MP *ja*. Therefore, the mere exposure to MP use in NS writing in the first half of the TC course did not trigger learner MP use, whereas the pedagogical intervention resulted in a drastic increase in MP use by the learners in approximation of the NS baseline on aggregate counts. By visually presenting the distribution of focal elements in the corpus subsets along the timeline, the plot tool is especially powerful for demonstrating the non-linear, explosive development of the target L2 feature in this study.[7]

7 The plot shows that the density of MP use by NSs also slightly increased toward the end of the correspondence (see Belz & Vyatkina, 2008 for a discussion). Although this was not the case in an earlier TC course iteration, learners' MP use also increased sharply after instruction (Belz & Vyatkina, 2005). Therefore, the MP increase in learner writing can be attributed to the instructional effect rather than to the increased exposure in the NS use.

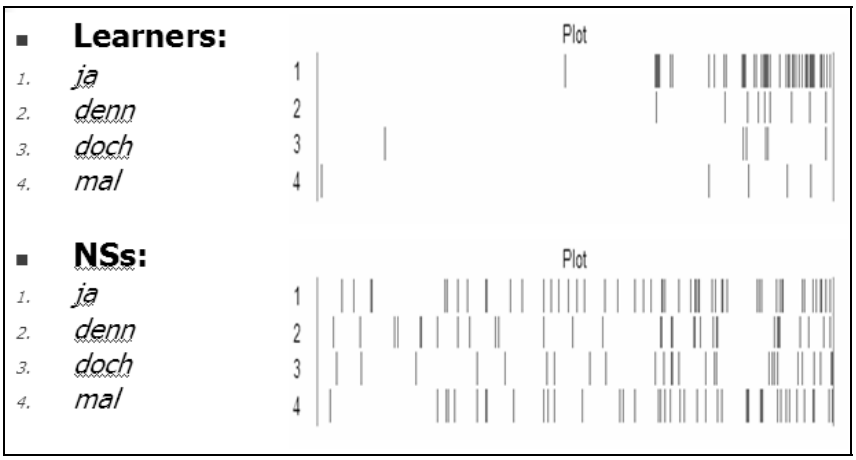

Figure 3. Dispersion plots of the MP use by learners and NSs

Wordlist

The wordlist-based analysis was performed for both CMC media types (email and chat) separately in order to investigate how common each focal MP is in comparison with other words in each respective subcorpus. For this purpose, the *WordSmith WordList* function was used. All running words from the corpus were retrieved in the form of a word list ordered by frequency. The most frequent word is ranked 1, the second frequent word 2 and so forth (Fig. 4). In addition, *WordList* provides summary statistics for each focal data subset. Table 1 presents such summary statistics for MP frequencies and the total number of German words broken down by medium (email and chat), time (pre-intervention and post-intervention), and participant group (learners and NSs).[8]

8 Email numbers declined after the first half of the course as email was mostly replaced by collaborative web-based projects in the second half.

Table 1. MP frequencies and total German words per medium, time, and participant group

medium	time	NSs		Learners	
		total words	MPs	total words	MPs
email	pre-intervention	14,685	54	10,284	3
	post-intervention	3,719	28	4,904	28
chat	pre-intervention	2,246	30	1,770	0
	post-intervention	3,421	52	3,631	43

First, the NS comparison baseline was explored. The word lists extracted from the NS chat and email subsets showed that MPs frequently appeared in NS writing, especially in synchronous CMC, i.e. chats. Among the total of 1,490 word types (or distinct words) contained in the NS chat subset, all four MPs cluster closely together and rank very high: from 47 to 57 in the frequency list. When normalized per 1000 words, the MP relative frequency ranges from 2.35 for *denn* to 3.54 for *ja*. Biber (2006) calls words that occur more than 1 time per thousand words of text highly frequent and considers them linguistic characteristics of the registers in which such words appear. Following this line of argument, MPs may be considered a linguistic characteristic of synchronous CMC discourse in German. The same holds for the NS email subcorpus, however to a lesser extent. The MP *ja* ranked 83 among 3,098 NS email word tokens and appeared on average 2 times per 1000 words and the MP *mal* ranked 132 and appeared 1.25 times per 1000 words, whereas the MPs *denn* and *doch* were still very frequent but appeared less frequently than 1 time per 1000 words. Next, word lists for the learner pre-intervention and post-intervention subcorpora were extracted. Whereas the pre-intervention chat list did not contain any MP instances, the post-intervention list contained all focal MPs. Notably, *ja* ranked 20th (Fig. 4), which places it among the most frequent words (mostly function words). In contrast, *mal, denn,* and *doch* were far below the NS baseline. A similar pattern was ascertained for emails. Therefore, the learners overused *ja* and underused three other particles in comparison with native speakers in both synchronous and asynchronous TC exchanges.

	Word	Freq.	%	Texts	%	RelC(3)
1	ICH	195	5.38	4	100.00	5.4
2	JA2	133	3.67	4	100.00	9.1
3	IST	121	3.34	4	100.00	12.4
4	DAS	81	2.24	4	100.00	14.6
5	WIR	64	1.77	4	100.00	16.4
6	UND	63	1.74	4	100.00	18.1
7	ES	57	1.57	3	75.00	19.7
8	NICHT	53	1.46	4	100.00	21.2
9	DIE	51	1.41	4	100.00	22.6
10	GUT	44	1.21	4	100.00	23.8
11	HABE	40	1.10	4	100.00	24.9
12	ABER	39	1.08	4	100.00	26.0
13	EIN	36	0.99	4	100.00	27.0
14	IN	35	0.97	4	100.00	27.9
15	SEHR	34	0.94	4	100.00	28.9
16	DU	33	0.91	4	100.00	29.8
17	EINE	33	0.91	4	100.00	30.7
18	AUCH	30	0.83	4	100.00	31.5
19	HABEN	30	0.83	4	100.00	32.4
20	JA1	30	0.83	4	100.00	33.2

Figure 4. WordList, learners, post-intervention chats

Keywords

The *KeyWords* tool allows automated comparison between two word lists. A key word is unusually frequent (or unusually infrequent) in comparison with the reference corpus (Scott, 2001). The statistic of log-likelihood is automatically calculated by this tool for marking significant and non-significant differences (Scott, 2001). Therefore, if the target feature appears among the key words while comparing two corpora, it is used in one of them with a significantly lower frequency (negative keyness) or higher frequency (positive keyness) in comparison with the other one. For this analysis, word lists for pre-intervention and post-intervention learner and NS subcorpora were extracted from *Telekorp* with email and chat data collapsed together to increase the frequency counts. Key word comparisons were conducted between the following pairs of subcorpora: 1) pre-intervention and post-intervention learner corpus; 2) pre-intervention

and post-intervention NS corpus; 3) pre-intervention learner corpus and NS corpus; 4) post-intervention learner corpus and NS corpus. The MP *ja* appeared as a key word for distinguishing between the learner pre-intervention and post-intervention corpus but no other MPs were key for this comparison (Fig. 5). Therefore, although learners increased their use of all MPs after the intervention, only the increase in their use of *ja* was significant. No MPs functioned as key words between the NS pre-intervention and post-intervention corpus, which means that the NS baseline did not change over time.

N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P
1	JA	1		50	0.58	-79.28	0.0000000000

Figure 5. MPs as key words between the pre-intervention and post-intervention learner corpus

The pre-intervention learner and NS corpus were distinguished by three MPs that appeared as key words: *ja*, *denn*, and *mal* (see Fig. 6). For the post-intervention learner-NS comparison, no MPs functioned as key words, which means that the learner MP use was not significantly different from NS use after the intervention. The frequency of the MP *doch* in both learner and NS CMC writing was too low to play a role in the differences found.

N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P
1	MAL	1		24	0.14	-34.62	0.0000000011
2	DENN	0		22	0.13	-38.44	0.0000000000
3	JA	1		29	0.17	-42.99	0.0000000000

Figure 6. MPs as key words between the pre-intervention learner and NS corpus

Concordance

Whereas *WordList* and *KeyWord* were used to compare frequencies of occurrence of separate words in different datasets, the next series of analyses was conducted to explore characteristic co-occurrence of patterns of words, or collocations (Firth, 1968; Sinclair, 1991; McEn-

ery, et al., 2006). The *WordSmith Concord* tool was used, which "locates all references to any given word or phrase within our corpus, showing them in standard concordance lines with the search word centered and a variable amount of context at either side" (Scott, 2001: 47, see Fig. 7–9). Having search words stacked and their left or right neighbors (collocates) alphabetized and highlighted facilitates comparison and finding regular patterns (O'Halloran & Coffin, 2004).

Collocations of German MPs have been thoroughly explored in NS oral discourse (e.g. Möllering 2004; Thurmair, 1991). These studies have shown that MPs frequently appear in collocation with the following linguistic structures: pronouns; modal verbs; other MPs; and formulaic expressions specific for each MP. Browsing the *Telekorp* concordance lines showed that the same linguistic structures were frequent neighbors of the focal MPs in both the NS and learner CMC writing. Following this preliminary finding, concordance lists for each MP and each data subset were further explored for specific contextual use.

The results of this analysis will be exemplified by collocations of *ja*, the most frequent MP in both the NS and learner discourse. It was found that, despite the similarity in the types of linguistic structures collocating with *ja*, there were also marked differences in the frequency and contextual meaning of these collocates. In terms of overall frequency, learners underused *ja* in collocations with pronouns, modal verbs, and other MPs and overused *ja* in formulaic patterns of appraisal in comparison with NSs. Moreover, a detailed contextual analysis of *ja* collocations with pronouns revealed additional fine-grained differences between the two populations.

NSs were found to use the MP *ja* in collocation with forms of the 2nd person personal and possessive pronouns *Du, ihr, dein* ('you', 'your') in 27.12% of all *ja* occurrences, and with forms of the 1st person personal and possessive pronouns *ich, mein* ('I', 'my') in 13.56% of all occurrences. In contrast, learners used *ja* in collocation with the 1st person pronouns with much higher frequency than with the 2nd

person pronouns: 33.33% of all *ja* occurrences with 'I', 'my' and 5.88% with 'you', 'your' (see Fig. 7, 8 for examples).

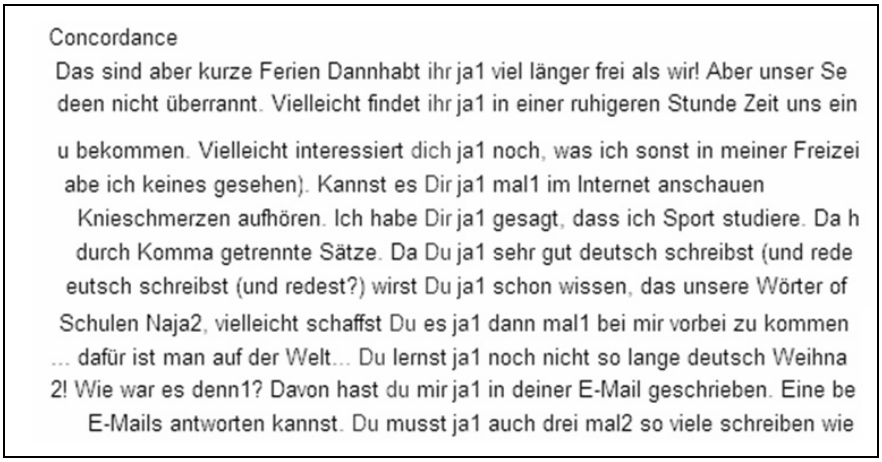

Figure 7. Collocations of *ja* with 2nd person pronouns, NSs

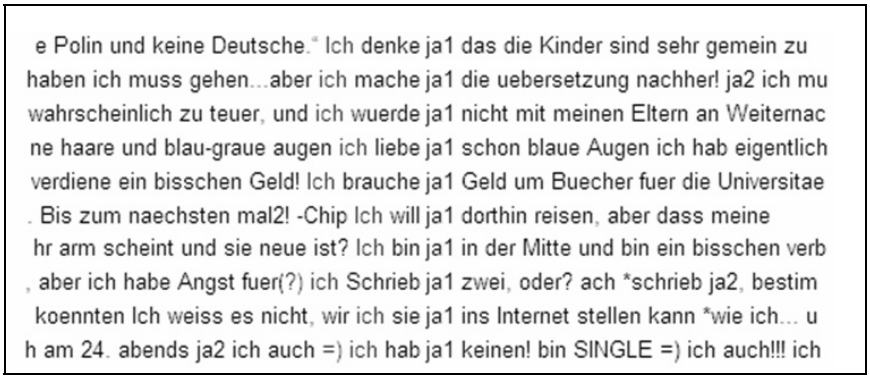

Figure 8. Collocations of *ja* with 1st person pronouns, learners

MPs also collocated with the plural form of the 1st person pronoun *wir* ('we') in both NS and learner data. However, the two populations used this pronoun differently. Waugh (2006) has shown that in spoken interactions personal pronouns may play an 'inclusive' interpersonal role, placing the speaker and the hearer in one and the same social group, or an 'exclusive' role, placing the hearer outside the speaker's social group. In *Telekorp* data, German NSs used *wir* fre-

quently and mostly inclusively, for referring to both German and American partners (Fig. 9). Americans, on the other hand, used *wir* only twice, both times exclusively, for referring to Americans as opposed to Germans.

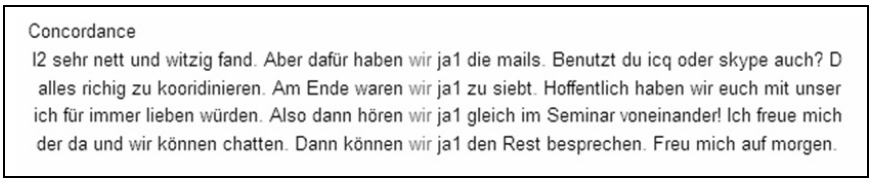

Figure 9. Collocations of *ja* with the inclusive pronoun *wir*, NSs

Based on this classification, personal and possessive pronoun instances in *Telekorp* can be grouped into speaker-oriented pronouns ('I', 'my', and exclusive 'we', 'our') and hearer-oriented pronouns ('you', 'your', and inclusive 'we', 'our'). After calculating frequencies for these categories, NSs were found to use the MP *ja* in collocation with hearer-oriented pronouns in 45.76% and with speaker-oriented pronouns in 15.25% of all occurrences. Learners, in contrast, were found to use *ja* in collocation with hearer-oriented pronouns in 5.88% and with speaker-oriented pronouns in 37.25% of all occurrences (Figure 10). Therefore, NSs use the MP *ja* with the highest frequency in hearer-oriented utterances, which is in line with the interpersonal MP meaning component "reference to shared knowledge" (Möllering, 2001, p. 132). In contrast, learners use the MP *ja* most frequently when they refer to themselves (as a person or as a group of Americans as opposed to Germans).

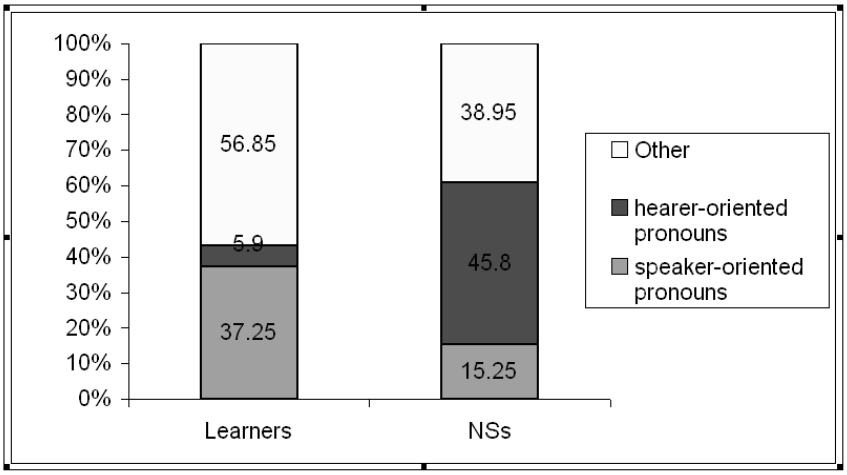

Figure 10. Proportion of collocations of *ja* with hearer-oriented and speaker-oriented pronouns, learners and NSs

Summary and Discussion

This chapter focused on applications of some corpus-based research methods and automated corpus analysis tools to TC discourse. First, the methodology of Contrastive Interlanguage Analysis, or CIA, was briefly reviewed and selected examples of CIA studies were listed. Next, advantages of combining CIA with research on telecollaboration were explained. Finally, examples from research that employed CIA for exploring a TC learner corpus were given, including the description of concrete applications of WordSmith Tools lexical analysis software. In this section, I summarize how a number of unique features of TC data facilitated a successful application of CIA methods and tools.

First, the *computer-mediated nature of the production data* allowed for automated saving of all of the TC discourse in computers. This allowed the researcher to explore the full course of L2 development of

language learners under study from the beginning to the end of the TC exchange. The text-only data did not require a transcription and could be used simultaneously with the TC course for research and teaching.

Second, a *rich variety of learner and task metadata* were collected and saved in the TC corpus. This allowed the creation of data subsets for analysis according to language (only German language data), medium (email – chat), date of production, etc.

Third, an *integrated NS corpus* compared learner productions to NS baseline which was valid because it was drawn from the very same interactions and not from an unspecified external NS corpus. In particular, contrasts were computed between learner and NS pre-intervention and post-intervention MP use. Word list and key words corpus analysis tools were used to demonstrate that learners drastically underused MPs in comparison to NSs before the intervention but there was no significant difference between the populations after the intervention. Moreover, after the intervention, the learners were found to use MPs with different frequency according to the CMC medium in the same fashion as the NSs (fewer MPs in emails than in chats). Finally, the most frequent MP in both NS and learner discourse was *ja*. This finding confirms the results of previous research (Cheon-Kostrzewa & Kostrzewa, 1997; Kasper & Rose, 2002) showing that the L2 acquisition path often leads from most frequent to less frequent language features.

Fourth, because *Telekorp* archived *dense developmental data* and not only pretest/posttest data, it was possible to include a developmental component into the analysis and compare learner post-intervention productions to their own pre-intervention productions as a baseline. Moreover, the *Plot* function of the *Concord* tool facilitated a comparison of the dispersion of individual MPs between learners and NSs. It was found that the learner development was rather uneven: not monotonic over the TC course but rather explosion-like after the intervention (Figure 3). Furthermore, the individual developmental paths of each participant can be tracked over time in addition to analysis of aggregate cohort data. Such analyses are beyond the scope

of this chapter, and the reader is referred to a series of studies which qualitatively and quantitatively analyze the development of individual participants from this study (Vyatkina, 2007; Belz & Vyatkina, 2008) and from another TC cohort (Belz & Vyatkina, 2005; Vyatkina & Belz, 2006).[9]

Fifth, applying CIA to TC discourse proved to be very efficient in demonstrating *L2 pragmatic development,* namely productive use of *interpersonal features,* as the result of pedagogical intervention. Fluency in using such features greatly helps learners develop advanced L2 competence with 'economical' means, as Weydt, Harden, Hentschel, and Rösler argue:

> Von einem gewissen Kenntnisstand des Deutschen an ist es viel ökonomischer, die immer wiederkehrenden Abtönungspartikeln zu lernen, als noch weiter an der Grammatik oder am übrigen Vokabular zu arbeiten. Man wird mit einem vergleichsweise geringen Aufwand an Arbeit ein erhebliches Mehr an Fähigkeit erreichen, idiomatisches Deutsch zu sprechen.[10.] (1983: 9)

This study confirmed the feasibility of explicit data-driven teaching of MPs. Despite ample exposure to the MPs in the NS discourse, only two learners used a total of 3 MPs at the pre-intervention stage. After the intervention, learners showed a dramatic increase in MP frequency and range in approximation of the NS baseline after 2–3 weeks of instruction. This result provides striking evidence in favor

9 It is worth noting that *all* participants made great strides in developing their production and/or awareness of German MPs although each of them followed his/her own developmental path. Some of them started using MPs right after the first instructional session while others waited until they became cognitively ready, and yet others did not start using MPs during the TC course but expressed a strong intention to use them in the future. These individual differences point to the limitations of aggregate quantitative studies and to the need to supplement them with qualitative case studies.

10 'From a certain level of knowledge of German onward, it is much more economical to learn the modal particles that always pop up than to work further on grammar or the rest of vocabulary. At a relatively low cost in work, one can attain a considerable enhancement in the ability to speak idiomatic German.' (Author's translation)

of explicit instruction in pragmatics, especially in comparison with untutored MP acquisition in study abroad settings that typically happens only after 20–30 months of exposure (Cheon-Kostrzewa & Kostrzewa 1997; Rost-Roth, 1999).

Sixth, the analysis of MP *collocations* showed that all patterns typical of NS data were found in the learner usage, however, some of these patterns were underused and others were overused. Although no significant aggregate frequency difference was found in the post-intervention data, there appears to be a rather large difference between NSs and learners with regard to MP contextual use. This analysis sheds light on the MP *semantic prosody* in NS and learner CMC writing, i.e. the contextual meaning "which is established through the proximity of a consistent series of collocates" (Louw, 2000: 57). In particular, it was found that NSs tended to use more "harmonic" MP/pronoun collocations and learners more "disharmonic" collocations (Aijmer, 2002: 68). In other words, collocations of MPs and hearer-oriented pronouns in NS discourse reinforced the interpersonal nature of both linguistic features, whereas the interpersonal nature of MPs conflicted with the speaker-oriented function of personal pronouns preferred by learners. This finding, again, lends support to explicit teaching of such fine-grained nuances of meaning with the help of corpus-based awareness-raising activities (see Vyatkina & Johnson, 2007 for pedagogical suggestions).

Conclusion

This chapter has demonstrated that TC discourse lends itself to explorations by corpus analysis methods and more studies in this direction are welcome and needed. Educators considering TC exchanges are encouraged to collect TC corpora and to include a research component in their planning process. First, they need to carefully follow the IRB procedures for protection of human subjects and agree these

procedures with their TC partners since they may differ from institution to institution and from country to country. Next, they should decide upon and carefully follow the procedure for saving, archiving, and organizing the TC data. Although using database software is advisable for archiving both primary data and metadata, one should at least devise a transparent system of folders for saving the data as separate text files. The annotation of corpus data will depend on particular research objectives.

Both corpus-driven and corpus-based (Tognini-Bonelli, 2001) studies may be conducted on TC corpora. The former are exploratory studies that involve browsing through raw corpus data, looking for potentially interesting features and patterns, and formulating hypotheses for future testing. The latter are studies that focus on testing hypotheses which may come from previous research conducted in non-TC settings or one's own exploratory corpus-driven studies. For such studies, problem-oriented corpus annotation for target features needs to be conducted prior to analysis.

TC corpus research designs can be particularly recommended for SLA researchers interested in developmental interventional studies exploring effects of TC corpus-based pedagogical interventions (Belz & Vyatkina, 2008), studies of hybrid media and genres, as well as studies of interactional linguistic and discourse features. It is worth mentioning that although this essay reported on the use of proprietary corpus software such as *FileMaker* and *WordSmith*, open source software should be highly recommended for future studies. An example of user-friendly open source software with incorporated corpus annotation and analysis tools is *UAM Corpus Tool*[11] (O'Donnell, 2011) and a number of other tools have been developed that are freely available for research purposes. Finally, the importance of using different lenses for learner corpus analysis should be especially pointed out. Corpus research is uniquely suited for combining quantitative and qualitative methods, aggregate and microgenetic analyses,

11 <http://www.wagsoft.com/CorpusTool/>.

and such combinations are necessary for providing thorough and multifaceted accounts of learner development.

References

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55–76). Amsterdam/Philadelphia: John Benjamins.

Aijmer, K. (2009). *Corpora and language teaching*. Amsterdam/Philadelphia: John Benjamins.

Aston, G., Bernardini, S., & Stewart, D. (2004). *Corpora and language learners*. Amsterdam/Philadelphia: John Benjamins.

Barlow, M. (2005). Computer-based analyses of learner language. In: R. Ellis & G. Barkhuizen (Eds.), *Analyzing learner language* (pp. 335–357). Oxford: Oxford University Press.

Belz, J. A. (2003a). From the special issue editor. *Language Learning & Technology, 7*(2), 2–5.

Belz, J. A. (2003b). Linguistic perspectives on the development of intercultural competence in telecollaboration. *Language Learning & Technology*, *7*(2), 68–117.

Belz, J. A. (2005a). Telecollaborative foreign language study: A personal overview of praxis and research. In D. Hiple & I. Thompson (Eds.), *Selected papers from the 2004 NFLRC symposium on distance education, distributed learning, and language instruction*. Honolulu: National Foreign Language Resource Center, University of Hawai'i. Downloaded 1 May 2011 from <http://www.nflrc.hawaii.edu/networks/nw44/belz.htm>.

Belz, J. A. (2005b). Corpus-driven characterizations of pronominal *da*-compound use by learners and native speakers of German. *Die Unterrichtspraxis/Teaching German, 38*(1), 43–59.

Belz, J. A. (2006). At the intersection of telecollaboration, learner corpus research, and L2 pragmatics: Considerations for language program direction. In J.A. Belz & S. L. Thorne (Eds.), *Internet-mediated intercultural foreign language education* (pp. 207–246). Boston, MA: Heinle and Heinle.

Belz, J. A. (2007). The role of computer mediation in the instruction and development of L2 pragmatic competence. *Annual Review of Applied Linguistics, 27*, 45–75.

Belz, J. A., & Kinginger, C. (2002). The cross-linguistic development of address form use in telecollaborative language study: Two case studies. *Canadian Modern Language Review 59*(2), 189–214.

Belz, J. A., & Kinginger, C. (2003). Discourse options and the development of pragmatic competence by classroom learners of German: The case of address forms. *Language Learning, 53*(4), 591–647.

Belz, J. A., & Reinhardt, J. (2004). Aspects of advanced foreign language proficiency: Internet-mediated German language play. *International Journal of Applied Linguistics, 14*(3), 324–362.

Belz, J. A., & Thorne, S. L. (Eds.) (2006). *Internet-mediated intercultural foreign language education*. Boston, MA: Heinle and Heinle.

Belz, J. A., & Vyatkina, N. (2005). Learner corpus analysis and the development of L2 pragmatic competence in networked intercultural language study: The case of German modal particles. *Canadian Modern Language Review, 62*(1), 17–48.

Belz, J. A., & Vyatkina, N. (2008). The pedagogical mediation of a developmental learner corpus for classroom-based language instruction. *Language Learning and Technology, 12*(3), 33–52.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Byrnes, H., & Sinicrope, C. (2008). Advancedness and the development of relativization in L2 German: A curriculum-based longitudinal study. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 109–138). New York: Routledge.

Cheon-Kostrzewa, B., & Kostrzewa, F. (1997). Der Erwerb der deutschen Modalpartikeln. Ergebnisse aus einer Longitudinalstudie (II). *Deutsch als Fremdsprache*, *3*, 150–155.

Dooly, M. (2008). Telecollaborative language learning: A guidebook to moderating intercultural collaboration and language learning. Bern: Peter Lang.

Ellis, N.C., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics—Introduction to the special issue. *Applied Linguistics, 27*(4), 558–589.

Firth, J. R. (1968). Linguistic analysis as a study of meaning. In F. R. Palmer (Ed.), *Selected papers of J. R. Firth 1952–59* (pp. 12–26). London: Longman.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes, 24*, 321–332.

Furstenberg, G., Levet, S., English, K., & Mallet, K. (2001). Giving a virtual voice to the silent language of culture: the Cultura project. *Language Learning and Technology, 5*(1), 55–102.

Ghadessy, M., Henry, A., & Roseberry, R. (Eds.) (2001). *Small corpus studies and ELT: Theory and practice.* Amsterdam/Philadelphia: John Benjamins.

Gilquin, G. (2008). Hesitation markers among EFL learners: Pragmatic deficiency or difference? In J. Romero-Trillo (Ed.), *Pragmatics and corpus linguistics: A mutualistic entente* (pp. 119–149). Berlin/New York: Mouton de Gruyter.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6*(4), 319–335.

Granger, S. (Ed.) (1998a). *Learner English on computer.* London: Longman.

Granger S. (1998b). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 145–160). Oxford: Oxford University Press.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam: John Benjamins.

Granger, S. (2003). The international corpus of learner english: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538–546.

Granger, S., Hung, J. & Petch-Tyson, S. (Eds.) (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Granger, S., & Meunier, F. (Eds.) (2008). *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins.

Guth, S., & Helm, F. (Eds.) (2010). *Telecollaboration 2.0. Language, literacies and intercultural learning in the 21st century*.. Bern: Peter Lang.

Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 77–116). Amsterdam: John Benjamins.

Kasper, G., & Rose, K. (2002). *Pragmatic development in a second language*. Malden, MA: Blackwell.

Kern, R. (2000). *Literacy and language teaching*. Oxford: Oxford University Press.

Kinginger, C., & Belz, J. A. (2005). Socio-cultural perspectives on pragmatic development in foreign language learning: Microgenetic case studies from telecollaboration and residence abroad. *Intercultural Pragmatics, 2*(4), 369–421.

König, E., & Requardt, S. (1991). A relevance-theoretic approach to the analysis of modal particles in German. *Multilingua, 10*(1/2), 63–77.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25*(1), 56–75.

Louw, B. (2000). Contextual prosodic theory: Bringing semantic prosodies to life. In C. Heffer, H. Sauntson & G. Fox (Eds.),

Words in context: A tribute to John Sinclair on his retirement (pp. 48–94). Birmingham, UK: University of Birmingham.

Lüdeling, A., Walter, M., Kroymann, E., & Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings from Corpus Linguistics 2005*, Birmingham. Downloaded 1 May 2011 from <http://www.corpus.bham.ac.uk/pclc/#corpora/>.

Mackey, A., & Gass, S. (2005). *Second language research. Methodology and design*. Mahwah, NJ: Lawrence Erlbaum Associates.

McCarthy, M. (1993). Spoken discourse markers in written text. In: J. Sinclair, M. Hoey & G. Fox (Eds.), *Techniques of description: Spoken and written discourse* (pp. 170–182). New York: Routledge.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Möllering, M. (2001). Teaching German modal particles: A corpus-based approach. *Language Learning & Technology*, *5*(3), 130–151.

Möllering, M. (2004). *The acquisition of German modal particles*. Bern: Peter Lang.

Myles, F. (2008). Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 58–72). New York: Routledge.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125–152). Amsterdam: John Benjamins.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578.

O'Donnell, M. (2011). UAM Corpus Tool [Computer software]. University of Madrid, Spain. Downloaded 12 November 2011 from <http://www.wagsoft.com/CorpusTool/index.html>.

O'Dowd, R. (2006). The use of videoconferencing and email as mediators of intercultural student ethnography. In J. A. Belz & S. L.

Thorne (Eds.), *Internet-mediated intercultural foreign language education* (pp. 86–120). Boston: Heinle and Heinle.

O'Dowd, R. (Ed.) (2007). *Online intercultural exchange: An introduction for foreign language teachers*. Clevedon, UK: Multilingual Matters.

O'Halloran, K., & Coffin, C. (2004). Checking overinterpretation and underinterpretation: Help from corpora in critical linguistics. In C. Coffin, A. Hewings & K. O'Halloran (Eds.), *Applying English grammar: Functional and corpus approaches* (pp. 275–297). London: Arnold.

Ortega, L., & Byrnes, H. (Eds.). (2008). *The longitudinal study of advanced L2 capacities*. New York: Routledge.

Paquot, M. (2008). Exemplification in learner writing: a cross-linguistic perspective. In S. Granger & F. Meunier (Eds.), *Phraseology in foreign language learning and teaching* (pp. 101–119). Amsterdam: John Benjamins.

Poos, D., & Simpson, R. (2002). Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. In R. Reppen, S.M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 3–23). Amsterdam/Philadelphia: John Benjamins.

Pravec, N. (2002). Survey of learner corpora. *ICAME Journal, 26*, 81–114.

Reder, S., Harris, K., & Setzler, K. (2003). A multimedia adult learner corpus. *TESOL Quarterly, 37*(3), 546–557.

Rogers, C. R. (1969). *Freedom to Learn*. Columbus, OH: Charles E. Merrill.

Rost-Roth, M. (1999). Der Erwerb der Modalpartikeln. Eine Fallstudie zum Partikelerwerb einer italienischen Deutschlernerin im Vergleich mit anderen Lernervarietaten. In N. Dittmar & A. Giacalone Ramat (Eds.), *Grammatik und Diskurs/Grammatica e discorso. Studi sull'acquisizione dell'italiano e del tedescho/Studien zum Erwerb des Deutschen und des Italienischen* (pp. 165–209). Tübingen, Germany: Stauffenburg.

Rundell, M., & Granger, S. (2007). From corpora to confidence. *English Teaching Professional, 50*, 15–18.

Sauro, S. (2009). Computer-mediated corrective feedback and the development of L2 grammar. *Language Learning and Technology, 13*(1), 96–120.

Schneider, J., & von der Emde, S. (2006). Dialogue, conflict, and intercultural learning in online collaborations between language learners and native speakers. In J. A. Belz & S. L. Thorne (Eds.), *Internet-mediated intercultural foreign language education* (pp. 178–206). Boston: Heinle and Heinle.

Scott, M. (2001). Comparing corpora and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs. In M. Ghadessy, A. Henry & R. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 47–67). Amsterdam/Philadelphia: John Benjamins.

Scott, M. (2008). WordSmith Tools (Version 5) [Computer software]. Liverpool, UK: Lexical Analysis Software.

Seidlhofer, B. (2002). Pedagogy and local learner corpora. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213–234). Amsterdam: John Benjamins.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (Ed.) (2004). *How to use corpora in language teaching*. Amsterdam: John Benjamins.

Thurmair, M. (1991). "Kombinieren Sie doch nur ruhig auch mal Modalpartikeln!": Combinatorial regularities for modal particles and their use as an instrument of analysis. *Multilingua, 10*(1/2), 19–42.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Tono, Y. (2000). A computer learner corpus based analysis of the acquisition order of English grammatical morphemes. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 123–132). Frankfurt am Main: Peter Lang.

Vyatkina, N. (2007). Development of second language pragmatic competence: the data-driven teaching of German modal particles

based on a learner corpus. The Pennsylvania State University, Unpublished doctoral dissertation.

Vyatkina, N., & Belz, J. A. (2006). A learner corpus-driven intervention for the development of L2 pragmatic competence. In K. Bardovi-Harlig, J. C. Félix-Brasdefer, & A. Omar (Eds.), *Pragmatics and language learning (Vol. 11)* (pp. 315–357). Honolulu: National Foreign Language Resource Center, University of Hawai'i.

Vyatkina, N., & Johnson, K. E. (2007). *Teaching German modal particles: A corpus based approach*. University Park, PA: CALPER Publications.

Warschauer, M. (Ed.) (1996). *Telecollaboration in foreign language learning*. Honolulu: Second Language Teaching and Curriculum Center, University of Hawai'i.

Waugh, L. (2006, September). Indices of identity and markers of ideological stance by proficient L2 speakers in their L1: Evidence from French conversational interaction. Lecture given at the Pennsylvania State University.

Weinrich, H. (1993). *Textgrammatik der deutschen Sprache*. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Weydt, H. (2006). What are particles good for? In K. Fischer (Ed.), *Approaches to discourse particles. Studies in pragmatics 1* (pp. 205–217). Amsterdam: Elsevier.

Weydt, H., Harden, T., Hentschel, H., & Rösler, D. (1983). *Kleine deutsche Partikellehre*. Stuttgart: Ernst Klett.